DSP543 Final Report

Jiayuan Zhang

May 10, 2019

Empirical investigation on peple’s purchasing behavior on black friday

Introduction

This report empirically tests whether people in different countries have different purchasing behavior on black Friday. The data comes from website kaggle. The original data set comes from a retail store that wants to study the consumers’ purchasing behavior, especially focus on the purchase amount. The dataset consists of 537577 observation. In this report, we respond to the appeal to test what will lead to the increase of purchase amount on Black Friday in this retail store. Two main methods, ANOVA and logistic regression, are applied to study the drivers of purchase amount.

The practical implication of the project is that we can provide insights on the consumers’ purchasing behavior by studying the data set. With a large size of data, the findings from the study are robusted. The restail store can use the results to boost the retails sales in the future black friday.

Data Description

Before we analyze on the data, we need to take a look on the data to find out the basic information. We plotted the graph to show the demographic information of the data. Figure 1 shows the Age distribution across the three cities. We find out that the three cities have similar distribution of the age group. All the cities show that age group 26-35 is the largest group that has the highest count. Group 18-25 and group 36-45 are also the major large groups. Figure 2 shows the gender group across the three cities. All three cities show that the retail store has more male than femal to purchase on Black Friday. Figure 3 shows that the occupation group across the three cities. From consumers who come from city A, occupation 0, occupation 4, and occupation 7 are the three groups that have larger population. From consumers who come from city B, occupation 0,4,and 7 are also the three groups that have largest population. The same thing happens in consumers from city C. Figure 4 shows that in three cities, consumers who are single have higher count than married group.

Figure 1 Count of age group

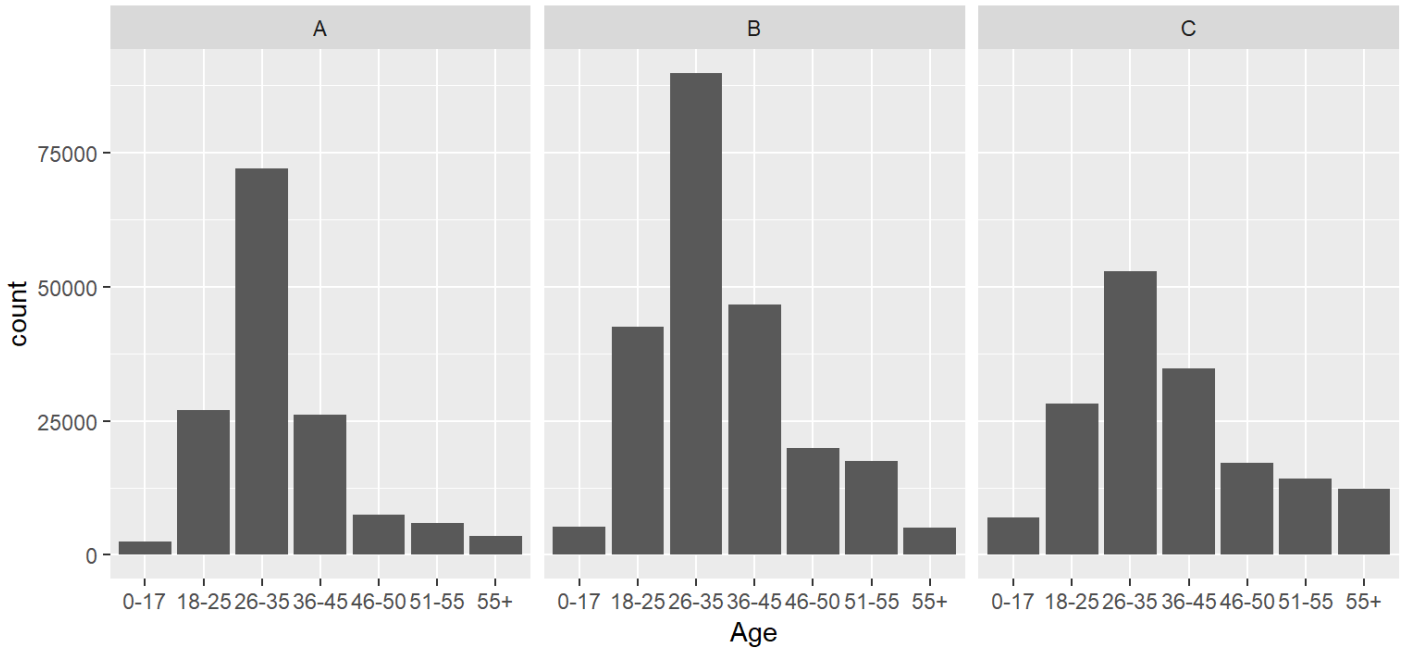


Figure 2 Count of gender group

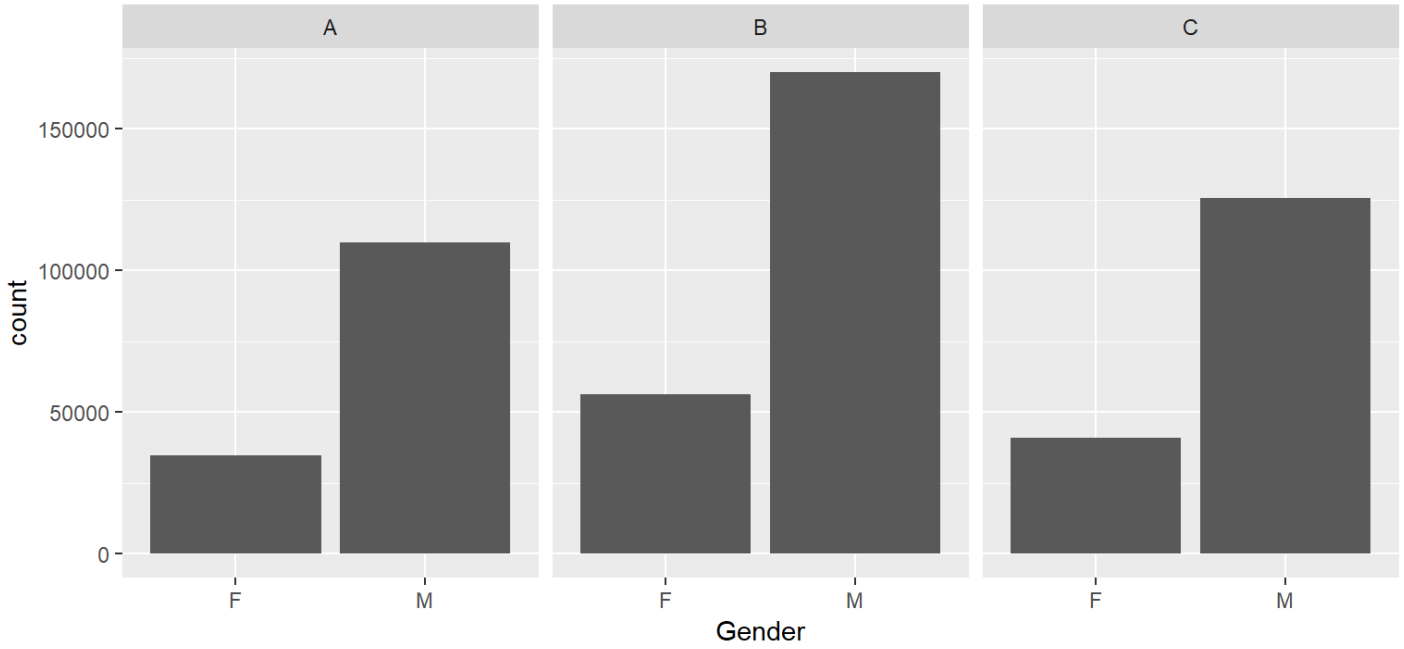


Figure 3 Count of occupation group

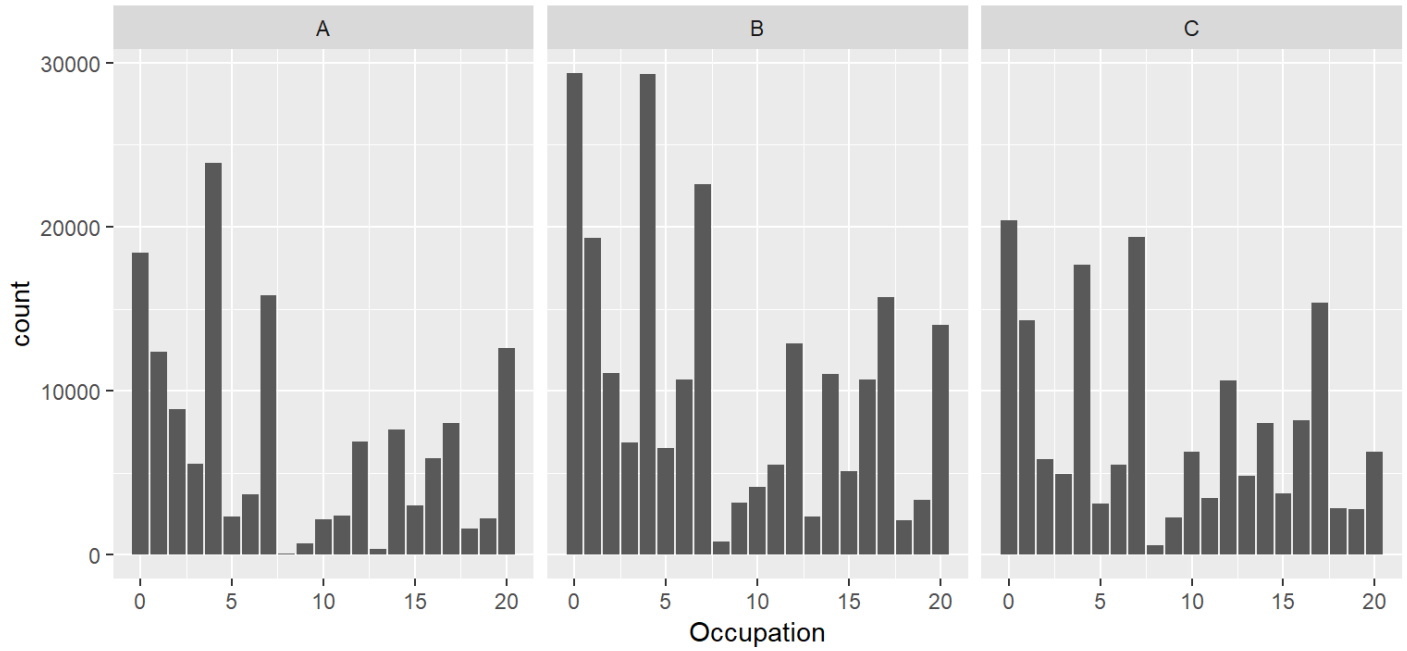
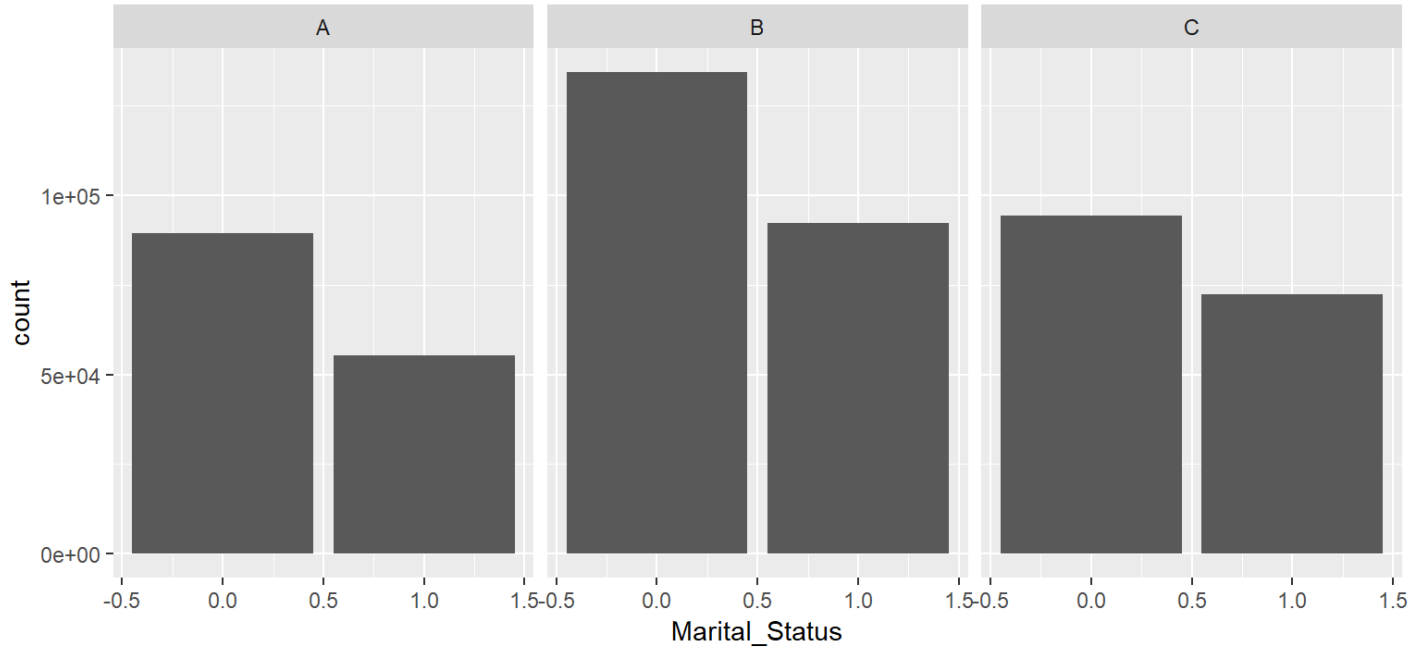
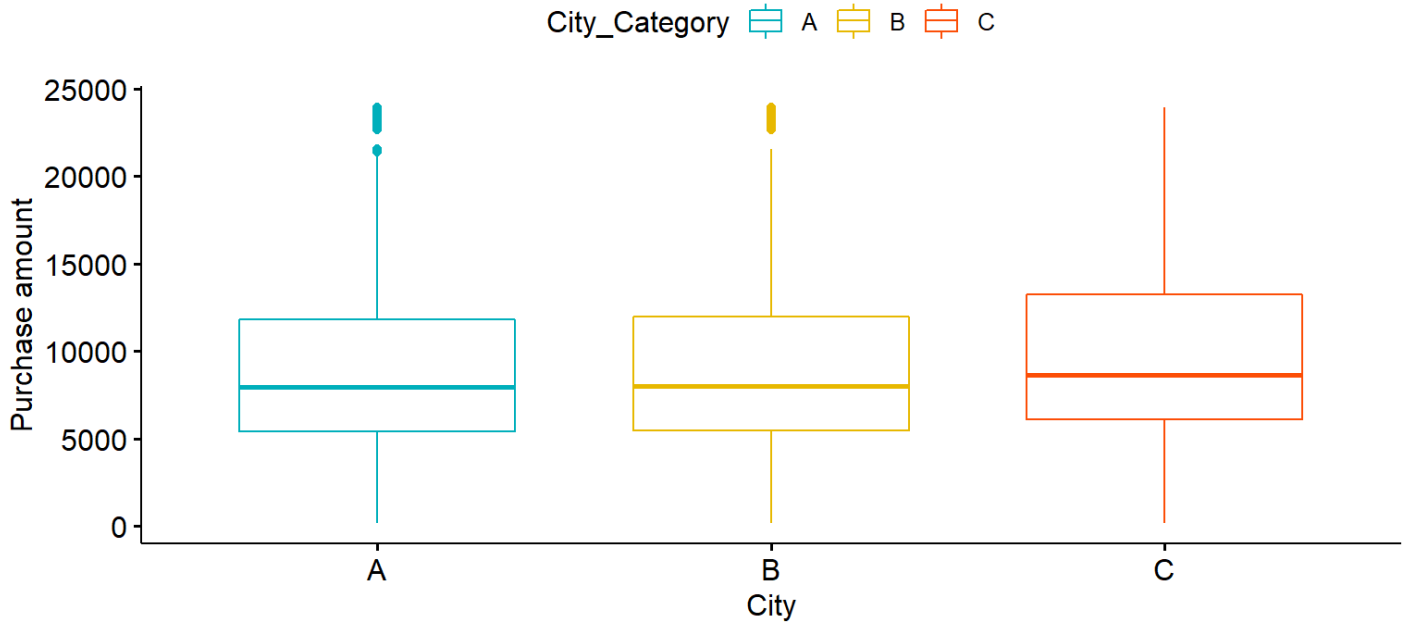


Figure 4 Count of marital status



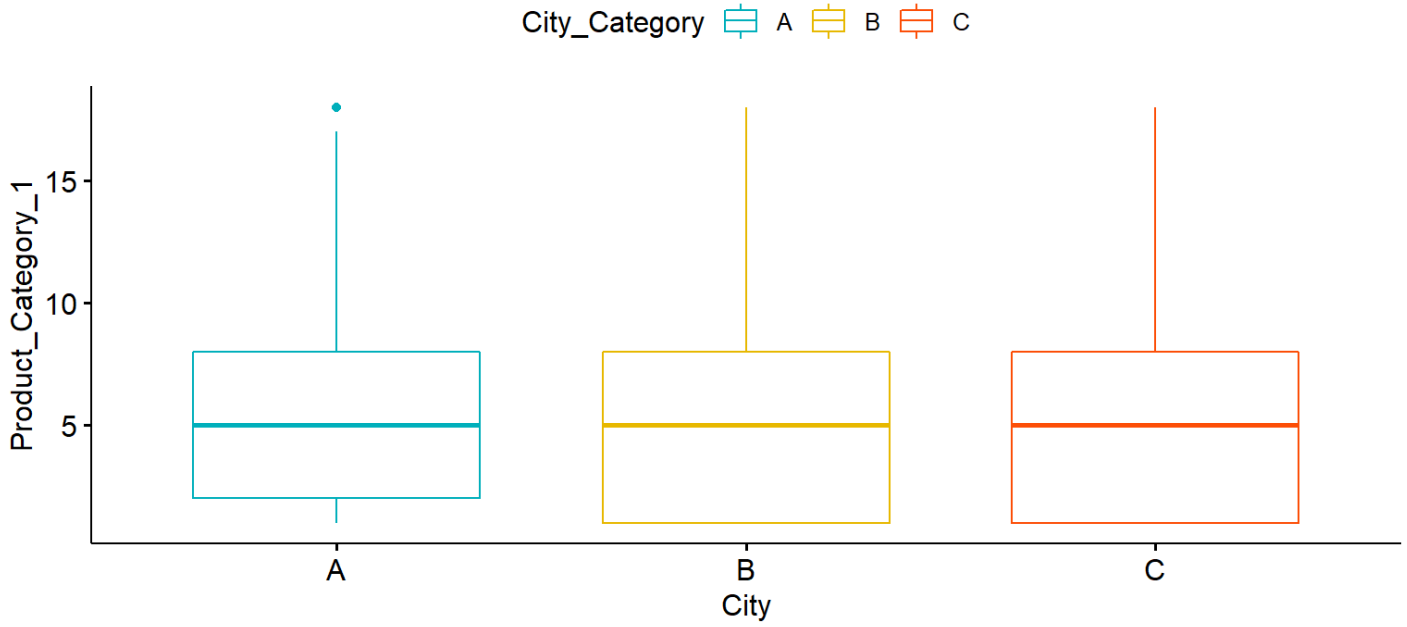
We then check the mean purcahse amount and standard deviation of three cities. The mean total purhcase amount of city A is 8958.0110137 and standard deviation is 4866.8965997. The mean total purhcase amount of city B is 9198.6578481 and standard deviation is 4927.0629648. The mean total purhcase amount of city C is 9844.441855 and standard deviation is 5109.4721004. Below we apply the boxplot to plot the mean of each city. City C has higher purchase amount than city B, which has higher purhcase amount than city A.

Figure 5 - Boxplot of purchase amount



Then we take a look at the purchase itmes of different categories. The mean purcahse amount of product category 1 of City A is 5.4370359 and standard deviation is 3.7278677. The mean purcahse amounot of product category 1 of City B is 5.3004199 and standard deviation is 3.7449832. The mean purcahse amounot of product category 1 of City C is 5.1659637 and standard deviation is 3.7736471. Below we plot the boxplot of the product category 1 in the three cities. The three cities share the similar purhcase amout of the product category 1.

Figure 6 - Boxplot of items in category 1

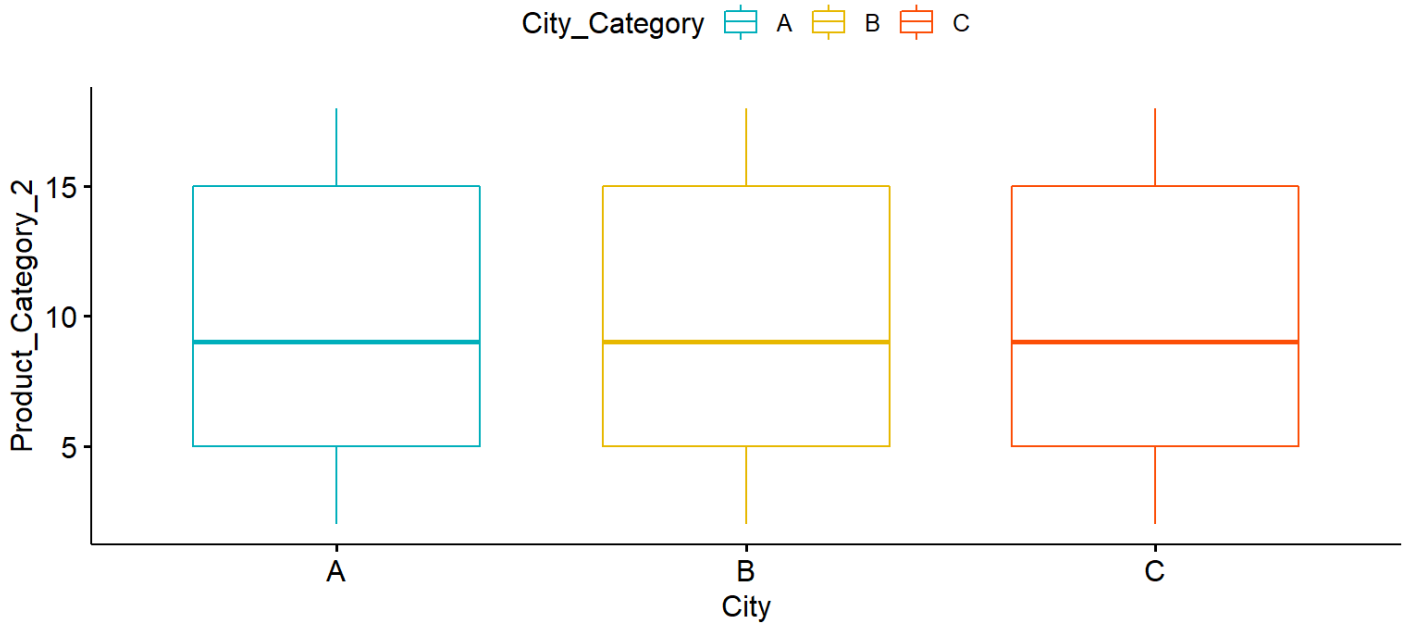


The mean purcahse items of product category 2 of City A is 9.9450261 and standard deviation is 5.0223497. The mean purcahse amounot of product category 2 of City B is 9.8257627 and standard deviation is 5.0779269. The mean purcahse amounot of product category 2 of City C is 9.7796489 and standard deviation is 5.1504693. Below we plot the boxplot of the product category 2 in the three cities.The three cities share the similar purhcase amout of the product category 2.

Figure 7 - Boxplot of items in category 2

The mean purcahse amounot of product category 3 of City A is 12.6825304 and standard deviation is 4.0957075.
The mean purcahse amounot of product category 3 of City B is 12.6748112 and standard deviation is 4.1111982.
The mean purcahse amounot of product category 3 of City C is 12.6543217 and standard deviation is 4.1614944.
The three cities share the similar purhcase amout of the product category 3.

Figure 8 - Boxplot of items in category 3



ANOVA analysis

In this section, we want to investigate whether different cities have different number of purchase amounts. Therefore, we run an ANOVA analysis on the purchase amount among the three cities. Below is the ANOVA results on the purchase amount among the three cities. The ANOVA results show that the three cities have different purchase amounts. We then go ahead to use Tukey test to see how the city is different from each city.The p-value of the comprison between city B and city A is 0, which means the purchase amount is different between city B and city A. The p-value of the comparsion between city C and city A is 0, which means the purchase amount

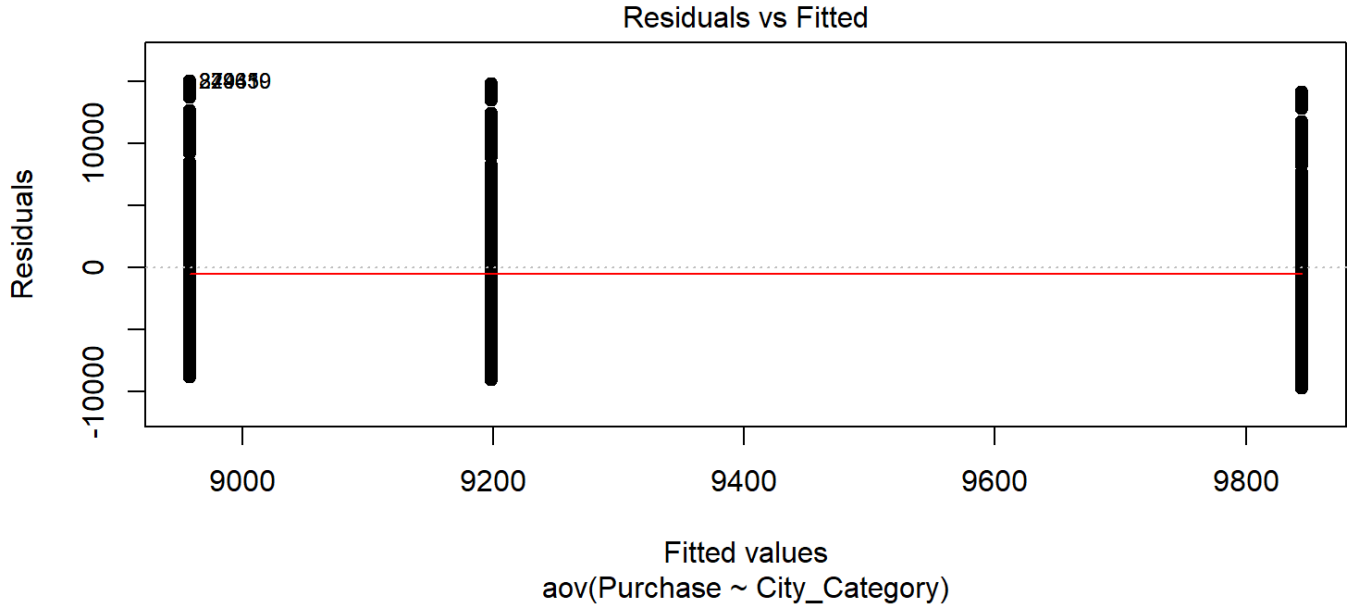
is different between city C and city A. The p-value of comparison between city C and city B is different, which means the purchase amount is different between city B and city C. Overall, the results show that the purhcase amount is different among the three cities.

```
## Call:
##    aov(formula = Purchase ~ City_Category, data = friday)
##
## Terms:
##                 City_Category    Residuals
## Sum of Squares   6.796357e+10 1.326961e+13
## Deg. of Freedom             2       537574
##
## Residual standard error: 4968.324
## Estimated effects may be unbalanced
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Purchase ~ City_Category, data = friday)
##
## $City_Category
##         diff      lwr      upr p adj
## B-A 240.6468 201.4540 279.8397     0
## C-A 886.4308 844.5734 928.2883     0
## C-B 645.7840 608.1907 683.3773     0
```

However, since we do not check the assumption of homogeneity and assumption of normality on the data, the ANOVA results might not be precise. Below we plot the graph to first check the homogeneity assumption on the data. From figure 9, we can see that the residuals are similar among the three cities. We then use bartlett test, which is to test whehter the residual is correlated with the fitted value, to test the assumption of homogeneity. We find that the p-value is less than 0, and it indicates that the residual is related with the fitted value. It shows that our assumption of homogeneity is violated.

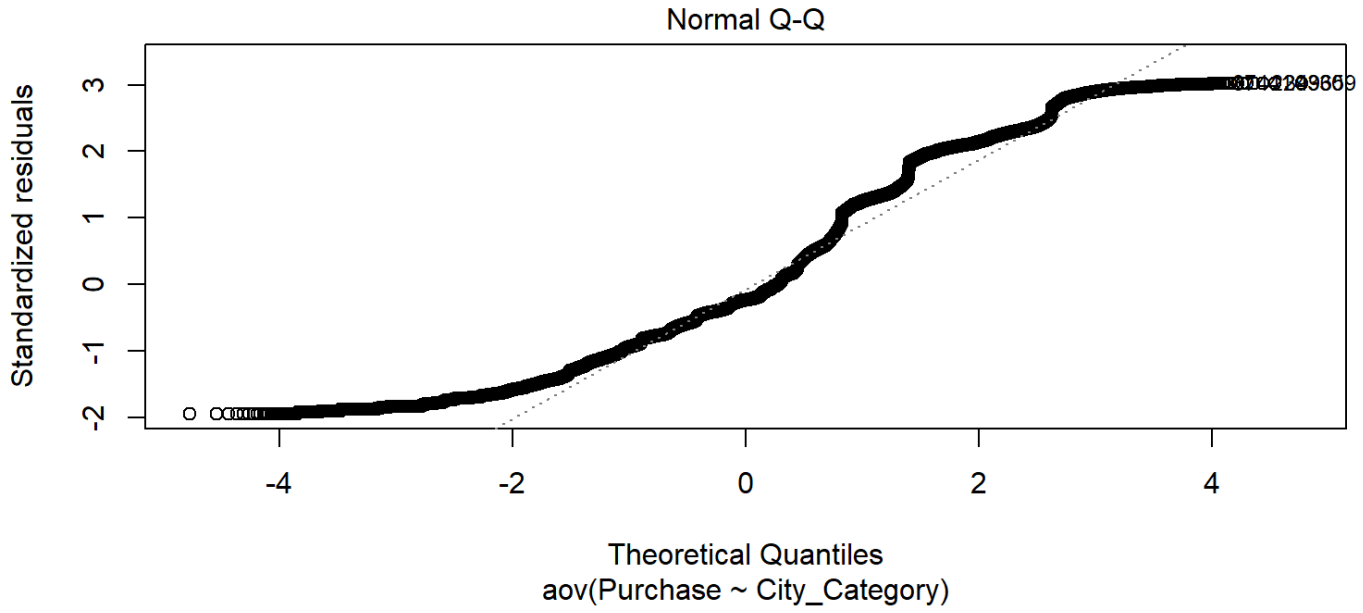
Figure 9 - check assumption of homogeneity



```
##
##  Bartlett test of homogeneity of variances
##
## data:  Purchase by City_Category
## Bartlett's K-squared = 418.91, df = 2, p-value < 2.2e-16
```

In the next step, we test the assumption of normality of the data. Figure 10 shows that relationship between the theoretical quantiles and standaradized residuals. If the assumption of normality is not violated, all the data will fall into the 45 degree line. However, figures 10 shows that not all the data fall to the 45 degree line, and most of them are not on the line. The graph shows that the assumption of normality is also violated in the data. Because both the assumption of homogeneity and assumption of normality are violated, we need to use the kruskal test, which is alternative test of ANOVA when these assumptions are not met. Below the kruskal test shows that the purchase amount of three cities are still different from each other.

Figure 10 check assumption of normality




```
##
##  Kruskal-Wallis rank sum test
##
## data:  Purchase by City_Category
## Kruskal-Wallis chi-squared = 2766, df = 2, p-value < 2.2e-16
```

We also run the ANOVA analysis on the age group and the amount of purchase. Below are the results that show comparision between different groups. The ANOVA results show that the purchase amount is different among different age groups. However, when we run a detail analysis on different age group, we have an interesting finding. The p-vale between age group 46-50 and 18-25 is not significant. It shows that people from age group 46-50 and people from age group 18-25 have the same purchase amount. The p-value between age group 46-50 and 26-35 is not significant either. It shows that these two groups do not have significantly different purchase amount. The p-value between age group 55+ and 36-45 is not significant, and it also shows that these two groups don't have different purchase amount. Since this is the data from large observation, the results show that younger adult and mid age group have the same purchase amount in this retail store.

```
##                Df    Sum Sq   Mean Sq F value Pr(>F)
## Age             6 6.471e+09 1.078e+09   43.49 <2e-16 ***
## Residuals   537570 1.333e+13 2.480e+07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Purchase ~ Age, data = friday)
##
## $Age
##                   diff        lwr       upr     p adj
## 18-25-0-17   215.07070   85.20433 344.93707 0.0000216
## 26-35-0-17   294.46209  169.31635 419.60783 0.0000000
## 36-45-0-17   381.35188  252.26796 510.43580 0.0000000
## 46-50-0-17   264.74540  125.10756 404.38324 0.0000005
## 51-55-0-17   600.48974  457.70399 743.27549 0.0000000
## 55+-0-17     433.77170  275.75253 591.79087 0.0000000
## 26-35-18-25   79.39140   22.71707 136.06572 0.0007180
## 36-45-18-25  166.28118  101.37215 231.19022 0.0000000
## 46-50-18-25   49.67470  -34.28511 133.63452 0.5859577
## 51-55-18-25  385.41905  296.32194 474.51615 0.0000000
## 55+-18-25    218.70100  106.80560 330.59641 0.0000002
## 36-45-26-35   86.88979   32.03211 141.74746 0.0000617
## 46-50-26-35  -29.71669 -106.17212  46.73873 0.9137885
## 51-55-26-35  306.02765  223.96380 388.09150 0.0000000
## 55+-26-35    139.30961   32.92933 245.68989 0.0021726
## 46-50-36-45 -116.60648 -199.35088 -33.86208 0.0006458
## 51-55-36-45  219.13786  131.18515 307.09057 0.0000000
## 55+-36-45     52.41982  -58.56652 163.40616 0.8061234
## 51-55-46-50  335.74434  232.92534 438.56334 0.0000000
## 55+-46-50    169.02630   45.92458 292.12803 0.0010140
## 55+-51-55   -166.71804 -293.37932 -40.05676 0.0020083
```

Linear regression

In this section, we will run a linear regression to see whehter different cities, different age group, and gender will impact the purchase amount. In the analysis, the dependent variable is the total purchase amount, which is continuous variable. The independent variables are the cities, age group, and gender, and these varialbes are categorical variables. Below is the results from the linear regression. The coefficient of city B is 245.8637438, and the p-value is 1.317155610^{-48}. It means that controlling all the other variables, consumers from city B will lead increase of purchase amount of 245.8637438 compared to consumers from city A. The coefficient of city C is 901.1602865, and the p-value is 0. It means that controlling all the other variables, consumers from city C will lead to increase of purchase amount of 901.1602865 compared to consumers from city A. The coefficient of gender is 689.8124658 with p-value 0 . It shows that controlling other variables, male will lead to increase of purchase amount 689.8124658 compared to female. Regarding the age group, the base level is the age group under age 18, all the coefficient of different age groups are positive, it shows that compared to age group under 18, all the age group will lead to the increase of purchase amount. It makes sense because age group under 18 is the group who does not have income. Among all the age groups, age group 51-55 has the highest coefficient 602.9235599, with significant p-value. The regression results show that city, age group, and gender will impact the purchase

amout. In addition to investigate the factors that impact the purchase amount, we also want to test whehter these factors will lead to higher probability of purchase amount over 10000. We will run the logistic regression in the section below.

```
##
## Call:
## glm(formula = Purchase ~ City_Category + Age + Gender, data = friday)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
##  -9890   -3448   -1204    2880   15797
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     8050.09      44.15 182.352  < 2e-16 ***
## City_CategoryB   245.86      16.78  14.653  < 2e-16 ***
## City_CategoryC   901.16      18.13  49.702  < 2e-16 ***
## Age18-25         298.60      43.97   6.791 1.11e-11 ***
## Age26-35         408.50      42.45   9.623  < 2e-16 ***
## Age36-45         433.20      43.67   9.919  < 2e-16 ***
## Age46-50         286.98      47.18   6.082 1.19e-09 ***
## Age51-55         602.92      48.26  12.493  < 2e-16 ***
## Age55+           283.58      53.42   5.308 1.11e-07 ***
## GenderM          689.81      15.73  43.867  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 24584010)
##
##     Null deviance: 1.3338e+13  on 537576  degrees of freedom
## Residual deviance: 1.3216e+13  on 537567  degrees of freedom
## AIC: 10673863
##
## Number of Fisher Scoring iterations: 2
```

Logistic regression

In this section, we will run a logistic regression to test whehter age, city, and gender will lead to higher chance of purchase amount over 10000. In the logistic regression model, the indepedent variables are still gender, age, and city. The dependent varialbe is a dummy variable with value “Yes” when purchase amount is greater than 10000 and value “No” when it is less than 10000. Below is the summary on the value “Yes” and “No” across different cities, gender, and age groups. We also plot the table of “Yes” and “No” across different cities, geneder, and age groups.

```
##      Age  City Gender   Yes    No
## 1   0-17 CityA      M   415   654
## 2  18-25 CityA      M  6957 13954
## 3  26-35 CityA      M 18249 36802
## 4  36-45 CityA      M  6453 12732
## 5  46-50 CityA      M  1763  4488
## 6  51-55 CityA      M  1639  2586
## 7    55+ CityA      M   921  2218
## 8   0-17 CityB      M  1243  2509
## 9  18-25 CityB      M 11373 19625
## 10 26-35 CityB      M 24944 43923
## 11 36-45 CityB      M 12571 23218
## 12 46-50 CityB      M  4971  8656
## 13 51-55 CityB      M  4883  8370
## 14   55+ CityB      M  1437  2276
## 15  0-17 CityC      M  1830  3103
## 16 18-25 CityC      M  8892 12776
## 17 26-35 CityC      M 17437 23987
## 18 36-45 CityC      M 10719 15386
## 19 46-50 CityC      M  4445  7347
## 20 51-55 CityC      M  4116  6390
## 21   55+ CityC      M  3216  5906
## 22  0-17 CityA      F   369  1059
## 23 18-25 CityA      F  1583  4531
## 24 26-35 CityA      F  4628 12369
## 25 36-45 CityA      F  2034  4923
## 26 46-50 CityA      F   357   859
## 27 51-55 CityA      F   453  1291
## 28   55+ CityA      F   118   233
## 29  0-17 CityB      F   483  1053
## 30 18-25 CityB      F  2922  8550
## 31 26-35 CityB      F  5706 15194
## 32 36-45 CityB      F  3260  7556
## 33 46-50 CityB      F  1759  4514
## 34 51-55 CityB      F  1148  3034
## 35   55+ CityB      F   336   979
## 36  0-17 CityC      F   579  1410
## 37 18-25 CityC      F  2110  4361
## 38 26-35 CityC      F  3666  7785
## 39 36-45 CityC      F  2942  5705
## 40 46-50 CityC      F  1712  3655
## 41 51-55 CityC      F  1185  2523
## 42   55+ CityC      F   921  2342
```

```
## , , City = CityA,  = Yes
##
##       Age
## Gender  0-17 18-25 26-35 36-45 46-50 51-55   55+
##      M   415  6957 18249  6453  1763  1639   921
##      F   369  1583  4628  2034   357   453   118
##
## , , City = CityB,  = Yes
##
##       Age
## Gender  0-17 18-25 26-35 36-45 46-50 51-55   55+
##      M  1243 11373 24944 12571  4971  4883  1437
##      F   483  2922  5706  3260  1759  1148   336
##
## , , City = CityC,  = Yes
##
##       Age
## Gender  0-17 18-25 26-35 36-45 46-50 51-55   55+
##      M  1830  8892 17437 10719  4445  4116  3216
##      F   579  2110  3666  2942  1712  1185   921
##
## , , City = CityA,  = No
##
##       Age
## Gender  0-17 18-25 26-35 36-45 46-50 51-55   55+
##      M   654 13954 36802 12732  4488  2586  2218
##      F  1059  4531 12369  4923   859  1291   233
##
## , , City = CityB,  = No
##
##       Age
## Gender  0-17 18-25 26-35 36-45 46-50 51-55   55+
##      M  2509 19625 43923 23218  8656  8370  2276
##      F  1053  8550 15194  7556  4514  3034   979
##
## , , City = CityC,  = No
##
##       Age
## Gender  0-17 18-25 26-35 36-45 46-50 51-55   55+
##      M  3103 12776 23987 15386  7347  6390  5906
##      F  1410  4361  7785  5705  3655  2523  2342
```

Below is the logistic regression of the results. We first analyze the gender. Becasue the base level is female, the coefficient 0.3490602 indicates that compared to female, male is 0.4177345 more likely than female to have purchase amount greater than 10000. Regarding the city group, the base level is city A. The coefficient of city B is 0.1078883, and it indicates that compared to people from city A, people from city B is 0.1139233 more likely than people from city A to have purchase amount over 10000. The coefficient of city C is 0.3041837, and it shows that compared to people from city A, people from city C is 0.355518 more likely than people from city A to have purchase amount over 10000. These logistic regression results are similar with the linear regression results. The base level of the age variable is age group below 18. All the coefficients of age group are positive, except for age group 55+. The coefficient of age group 55+ is -0.0660788, the negative value shows that compared to age group below 18, age group over 55 are 0.0639429 less likely to have purchase amount over 10000. It is interesting to

notice that there is a difference on age group 55+ from the two regression results. The linear regression shows that compared to age group below 18, age 55+ will lead to increase of purchase amount.However, the logistic regression shows that compared to age group above 55, age group below 18 is more likely to have purchase amount over 10000! This insightful result shows that the retail store can increase purchase amout if they target more on age group below 18. Among all the group ages, age group 36-45 has the highest coefficient 0.0911867, and it shows that age group 36-45 is 0.0954735 more likely than age group below 18 to have purchase amount over 10000. This logistic regression result is consistent with the linear regression result.

```
##
## Call:
## glm(formula = cbind(Yes, No) ~ F1 + F2 + F3, family = binomial,
##     data = dat1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -6.7978  -1.7084   0.0626   2.8174   7.6027
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.109727   0.019087 -58.140  < 2e-16 ***
## F118-25      0.072178   0.018868   3.826 0.000131 ***
## F126-35      0.085793   0.018227   4.707 2.51e-06 ***
## F136-45      0.091187   0.018733   4.868 1.13e-06 ***
## F146-50      0.013568   0.020256   0.670 0.502981
## F151-55      0.090615   0.020642   4.390 1.13e-05 ***
## F155+       -0.066079   0.022951  -2.879 0.003987 **
## F2CityB      0.107888   0.007236  14.909  < 2e-16 ***
## F2CityC      0.304184   0.007719  39.405  < 2e-16 ***
## F3M          0.349060   0.006904  50.560  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4858.4  on 41  degrees of freedom
## Residual deviance:  488.5  on 32  degrees of freedom
## AIC: 893.75
##
## Number of Fisher Scoring iterations: 3
```

Conclusion

This report analyzes the purchaseing behavior on consumers from a retail store. Overall, consumers from city C have the highest purchase amount among consumers from all the cities. The retail store can think about specific ways to adverties on consumers from city A and city B. Consumers from city A contribute less to purchase amount and have least chance to have purchase amount over 10000. There migh be some factors lead to this result. If the retail store can find out the reasons behind, it can lead to the increase of the sales on black friday. In addition, this study also finds that male lead to the increase of purchase amount and have higher chance than female to have purhcase amount over 10000. This finding is surprising. The retail store can focus on male consumers to boost sale on black friday. The results also show that age group 36-45 is the group that has the highest purchase amount to total purchase amount and has the highest probability to have purchase amount over 1000 compared to age group below 18. The retail store can also focus advertising on these groups to maintain the high sales.